NOTES AND LITERATURE

BIOMETRICS

Recent Contributions to Theory.—For obviously necessary reasons biometric work in the past has fallen, and for a considerable time in the future will continue to fall into two sharply separated categories. The first of these is the development of mathematico-statistical methods for dealing with biological material. The second is the concrete application of these methods to biological problems. Developments of higher mathematical theory will never be in and for themselves of particular interest to the great majority of biologists. Their only interest to the biologist exists in what they are practically good for; if there is developed any new mode of thinking or of investigation which will help to solve a problem it is of importance to know about it so that use may be made of it. The same consideration applies in regard to the applications of biometrical theory. To the biologist who is not cultivating this particular field of research the results obtained in it will be of interest only in so far as they appear to make significant contributions, expressible in terms intelligible to any non-mathematically trained person, to the solution of significant problems. There can be no doubt that valuable contributions to biology are being made by workers in biometry. Further, it is believed that this fact would be more generally recognized than is now the case, were the methods and results stated in a terminology less strange and repellent to the average biologist than is the mathematical. In accordance with this belief it will be the aim of these notes to set forth as clearly as may be and in a non-technical way the significant advances in biometrical research. Recent contributions to biometric theory will be considered in the present note. Work in the direction of the application of biometric methods to the solution of concrete biological problems will be discussed in subsequent notes.

The degree of trustworthiness of any reasoning from statistical constants of whatever sort must ultimately depend upon the values of the "probable errors" of these constants. The probable error is a measure of the limits of accuracy of the constant to which it applies. A number of recent papers deal

with various points regarding the general theory of probable errors. W. F. Sheppard, in a paper giving the mathematical proof of his corrections for the moments of frequency distributions,[1] points out that theoretically the raw value of the second moment rather than the value obtained by applying his corrections should always be used in calculating the probable errors of the mean and standard deviation. The present writer[2] shows by the consideration of concrete examples that the point made by Sheppard is not likely ever to be of any *practical* significance to the working biometrician. It makes no sensible difference in the actual result whether one does or does not use the corrected value of the second moment in determining these probable errors. In the same paper it is shown that a very considerable error may be made by using the formula ordinarily given for the probable error of the standard deviation when the distribution of variation differs considerably from the normal curve in certain particulars. One must always be cautious in assuming that a deviation from normality will be of no consequence in calculating probable errors.

A novel and important point regarding the probable error of a mean has been thoroughly investigated and solved by that mysterious person "Student,"[3] who has previously contributed to the pages of *Biometrika* under the same *nom de plume*. In passing it may be remarked that the propriety of publishing serious contributions to science under an assumed name seems questionable; one had supposed that those conditions no longer existed which in earlier times made such a course not merely desirable, but often indeed necessary if the same individual were to continue to contribute to science. To return to the point. In all kinds of experimental science the following type of problem very frequently arises; a series of say 10 or fewer experiments are tried under certain experimental conditions which may be denoted by the letters *a, b, c, d, . . . n*. A second series of equal extent is then tried with, say, one experimental condition changed so that the conditions are now *q, b, c, d, . . . n*. With what degree of probability may it be asserted that an observed differ-

[1] Sheppard, W. F. The Calculation of the Moments of a Frequency Distribution. *Biometrika*, Vol. V, pp. 450–459, 1907.

[2] Pearl, R. On Certain Points concerning the Probable Error of the Standard Deviation. *Ibid.*, Vol. VI, pp. 112–117, 1908.

[3] "Student." The Probable Error of a Mean. *Ibid.*, Vol. VI, pp. 1–25, 1908.

ence in the results of the parallel sets of experiments is *real*, that is, is due to the changing of condition a to q and not merely a chance result of random sampling? On account of the statistical smallness of such experimental series, coupled with the fact that the results usually deviate widely from the normal distribution of errors, the usual theory of the probable error of a class frequency or of a mean fails to give an adequate answer to the problem. "Student" has done the great service of providing formulas to cover the case. He furthermore gives a table of probabilities for series involving four to ten experiments, inclusive, wherefrom the answer to such a question as that propounded above may be read off at once when the experimental data are at hand. In the original paper concrete illustrations of the use of the table are given, based on published experimental data in the fields of physiology and of agriculture. No particular mathematical knowledge or skill is necessary to use this table and its importance to the experimental worker is obvious.

On this same topic of probable error Pearson and Lee[4] have recently made an important contribution. While the investigation itself is of a complex mathematical character, the essential point is this: Suppose a number of bodily characters (say six to eight) to be measured in a large sample of a population. There are likely to be a very few individuals out of the whole sample that differ rather widely from all the other individuals. What is the probability that such outlying individuals do not really (*i. e.*, biologically or genetically) belong to the population from which the sample is drawn, but only happen to get into the sample by accident? The solution to this problem is what Pearson and Lee attempt to give.

Along with these investigations on probable errors is to be included a recent paper of Pearson's[5] on the relation of past experience to future expectation. The nature of the problem discussed may be indicated by an example. Suppose that a sample of 100 individuals out of a population be examined and 2 per cent. of these individuals be found to be suffering from some particular disease. What percentage of a second sample of 100 may reasonably be expected to show the disease? The

[4] Pearson, K., and Lee, A. On the Generalized Probable Error in Multiple Normal Correlation. *Ibid.*, Vol. VI, pp. 59–68, 1908.

[5] Pearson, K. On the Influence of Past Experience on Future Expectation. *Phil. Mag.*, 1907, pp. 365–378.

answer usually given is that 2 per cent. is to be expected, with a probable error of a certain amount. In calculating this probable error in the ordinary way it is assumed that the distribution of chances is given by the normal or Gaussian curve of errors. Pearson shows that this is not true except (1) when the first sample is indefinitely large in proportion to the second, and (2) when the characteristic does not occur in either a very large or very small percentage of the population. He then proceeds to develop general formulæ from which one may determine first the average expectancy for the second sample and its true probable error, and second the frequency of future samples having varying degrees of the characteristic under consideration. These methods are fully illustrated with numerical examples. The usefulness of the methods for biological work may be indicated by an example. Suppose that in a sample of 200 tadpoles, 106 are males and 94 females. Suppose, further, that in a second sample of 300 collected later in the season the proportion is 171 males to 129 females. Is it to be concluded that between the collections some factor has come into operation tending to the greater production of males? Or, on the other hand, is the proportion exhibited in the second sample what might be regarded as a reasonable chance deviation due to random sampling in a population wherein the sex, determining factors had not changed since the first sample was collected? The paper under discussion furnishes the methods whereby questions of this sort may be definitely answered.

Another recent contribution of Pearson's[6] furnishes new methods of determining the degree of correlation between variates. The methods are adapted for use in cases where for one reason or another it is not feasible to use the ordinary product-moment method of finding a correlation coefficient. It is neither possible nor desirable to enter upon any detailed discussion of the mathematical features of these new methods here. Three new methods are given. The first, or "difference method" of determining correlation, furnishes a very simple and tolerably accurate way of deducing a correlation coefficient from a symmetrical (*e. g.*, homotypic) correlation table. The other two methods are con-

[6] Pearson, K. Mathematical Contributions to the Theory of Evolution —XVI. On Further Methods of Determining Correlation.

Draper's Company Research Memoirs. Biometric Series, IV, pp. 39, 1907. (London: Dulau & Co.)

cerned with the problem of determining correlation from material arranged in ranks. The most obviously simple way of dealing with any mass of statistical material is to arrange the individuals in a long row in order of ascending magnitude of some character. It is not even necessary that the character be accurately measurable on a quantitative scale to make such a row. The second, and on the whole most important of Pearson's new methods, shows how to deduce true variate correlation from the correlation of ranks. The method is less accurate than the usual product moment method. The third method deals with the determination of variate correlation from the positive difference of ranks. Regarding these last two methods Pearson's general conclusion is "that variate correlations found by ranks may prove to be a useful *auxiliary* method of dealing with correlation, when it is needful to give a rough answer to a problem in a brief time, or when the material itself is incapable of being accurately measured."

In connection with this matter of the analysis of statistics by the method of ranks or grades it is of interest to note that Galton has recently published a table[7] of deviates of the normal curve which will be found useful in practical statistical work. In the text accompanying the table (which is calculated by W. F. Sheppard) the use of the method of grades in dealing with material not susceptible of exact numerical measurement is discussed.

Of value for practical biometrical computing work is an editorial note[8] in the current number of *Biometrika* giving an easily calculated and extremely exact formula for the Γ function. This function is frequently used in statistical work, particularly in curve fitting.

RAYMOND PEARL.

PROTOZOA

Some Amœba Studies.—The amœboid organisms include a vast number of minute bits of glairy protoplasm which live in widely different places sometimes free in ponds and pools, sometimes confined as parasites, to the narrow limits of a single tissue. To

[7] Galton, F. Grades and Deviates. *Biometrika*, Vol. V, pp. 440–406, 1907.

[8] On a Formula for Determining $\Gamma(x+1)$. *Ibid.*, Vol. VI, pp. 118–119, 1908.